# Multiple testing on Experimental Economics Data

**Dongyin Hu**
Department of ECE
A59006035

## 1 Introduction

Benefited from the improvements in storage, communication and new technologies, the size of the data people have access to nowadays is ever growing. As reported in [2], the medical image research field has observed an growth of the geometric mean dataset size in an exponential trend. The researchers are more and more likely to deal with large dataset, and as well conducting many hypothesis tests to discover insights of the data. However, suppose each test is conducted at level $\alpha$ and the $p$-values are independent, then the family-wise error rate (FWER), the probability of one or more false rejections, equals $1-(1-\alpha)^N$ and grows exponentially as the number of null hypotheses increases. We are $99.4\%$ likely to make a false rejection when the size of total testing is $N = 100$ and $\alpha = 0.05$. This is the multiple testing problem.

In this project, a sequence of works ([1] and [3]) is followed to understanding how a FWER control procedure may affect the conclusion derived from the same dataset. The proposal is organized as follows: section 2 offers the background information about the two mentioned works and summarizes the procedure of data collection; section 3 shows a exploratory data analysis on the dataset; and section 4 replicates several conclusions mentioned in [1] in a different way to serve as the baseline for comparing with the results obtained form the proposed FWER procedure.

The codes for the project are uploaded at here.

## 2 Background

The work [1] investigates a charitable fundraising problem. To increase the revenue, using a scheme called matching gift is common in practice. A matching gift is a gift that is committed by a donor(s) conditioned on the contributions of others at a given rate, up to a maximum amount this leadership donor is prepared to give. A matching rate of $X : 1$ means that for each dollar contributed by others, the leadership donates $X$ dollars. It is empirically believed that a higher matching rate have noticeable power to influence future contributions.

To examine this rule of thumb, researchers sent direct mail solicitations to 50,083 prior donors who have given to the Organization at least once since 1991. A number of 16,687 subjects, or 33% of the sample are randomly assigned as the control group. All donors receive a four-page letter identical in all respects except the treatment letters included an additional information on the matching offer. The match offer is randomized with equal probability in three aspects:

1. The matching ratio ($1 : 1$, $2 : 1$, or $3 : 1$),
2. The maximum amount of the matching offer ($25,000, $50,000, $100,000 or unstated),
3. the suggested donation amount (equals, 1.25 times or 1.5 times the highest previous donation).

The response variables are 1) whether the donor contributed within a month after receiving the letter, 2) the dollars given without the matching amount (raw contribution), 3) the dollars given with the matching amount ($X + 1$ times the raw contribution), and 4) the change in the amount given. The

data are analyzed using Probit model (on variable 1) and regression model on the other variables, and some of the conclusions are:

1. Using matching grants increases both the revenue per solicitation and the probability that an individual donates;
2. Larger match ratios (i.e., 3 :1 and 2 :1) relative to smaller match ratios (1 :1) have no additional impact;
3. The matching grant works for donors in red states but not blue states;

A later work [3] revisited the conclusions made in [1] by analyzing the data from a perspective of multiple testing. For each unit (sample), there are multiple subgroups (for example, residing in a red or blue state, and gender) and multiple outcomes attached with it. A procedure with asymptotic FWER control and balance is proposed. Mathematically, the procedure ensures that:

$$\limsup_{n\to\infty} \mathrm{FWER}_Q \leq \alpha \tag{1}$$

$$\lim_{n\to\infty} Q\{\text{reject } H_s\} = \lim_{n\to\infty} Q\{\text{reject } H_{s'}\},\ \forall s, s' \in S_0(Q), \tag{2}$$

where $Q$ is the distribution of data, $\mathrm{FWER}_Q$ is the FWER with respect to Q, $S_0(Q)$ is the index set of all null hypotheses that are true under the distribution $Q$, and $s$, $s'$ are indices of null hypotheses. By applying the FWER procedure, the significance for different null hypotheses changes.

## 3 Exploratory Data Analysis

In this section some basic statistics about the data is explored. We define the treatment of interest as the matching ratio, the outcomes of interest are all four outcomes, and the subgroups of interest are four combinations of residing in a red/blue state and red/blue county, following the testings done in [3]. More aspects may be examine if time permits.

The original data is available here and a updated version is available here. As noted by the author, the latest one includes a single, merged dataset, and the latest dataset contains more metadata, for instance, variable labels.

### 3.1 Summary statistics

The summary statistics is listed as in table 1. Some subgroups more than the interested subgroups are listed as well for better understanding of the dataset. Many statistics are similar as in [1] but table 1 further offers a summary on household size and median income information, as brought up during the presentation.

### 3.2 Correlation between variables

A correlation matrix is calculated for a selected subset of variables as shown in figure 1. The dollars given with/without matching amount (second and third rows/columns) are highly correlated with the first variable `gave`, which servers as a sanity check. However, little correlation is found with respect to other variables. A interesting finding is that the median household income is slightly negatively correlated with the variables indicating residing in red state or county.

### 3.3 Distribution of interested outcomes

Figure 2 summarizes the empirical cumulative probability function (eCDF) for the latter three dependent variables. Responding or not is a binary variable so its eCDF is omitted. The x-axis of figures 2a and 2b are shown in log scale since most of the values are 0. Figure 2c is shown in the normal scale.

From the distribution, we can observe that all of the distributions seems to be different between the treatment group and the control group. Yet note that the first two figures are in log scale, so it the difference is amplified.

Figure further breaks down the treatment groups based on the matching ratios, and shows the mean of all four outcomes. The trend observed from eCDF seems consistent with the mean comparison.

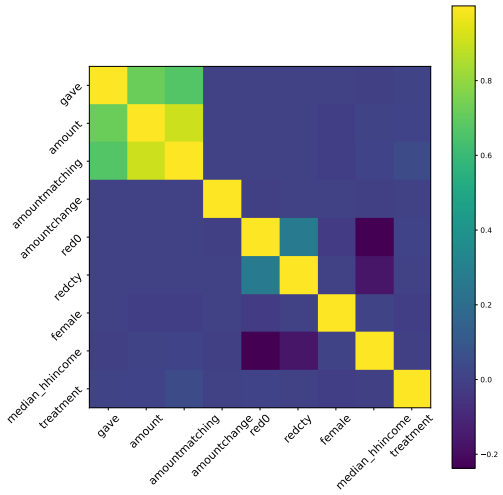



Figure 1: The correlation matrix for a subset of variables. The first four labels on x/y-axis are the four outcomes; `red0` and `redcty` represents whether the donor resides in a red state or county; `female` represents the gender; `median_hhincome` is the median of household income; `treatment` represents the treatment status.

## 4 Initial Analysis

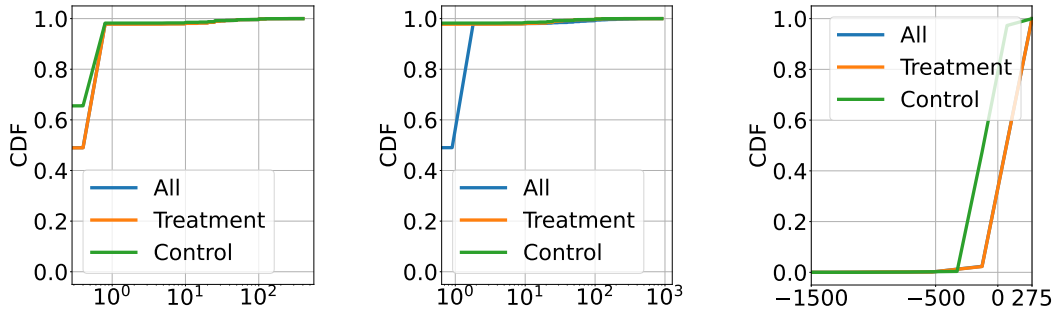
Following the work [1] and [3], in this section the conclusions listed in section 2 are tested again using the Wald test and permutation test. These testings are *different* from the methods used in these two papers.

In a general manner, we denote indices set of null hypotheses as in work [3]:

$$S = \{(d, d', z, k),\ d, d' \in D, z \in Z, k \in K\}\,, \tag{3}$$

| Subgroup | All | Treatment | Control |
|---|---|---|---|
| **Member Activities** | | | |
| # of previous donations | 8.04(11.39) | 8.04(11.39) | 8.05(11.40) |
| # years since initial donation | 6.10(5.50) | 6.08(5.44) | 6.14(5.63) |
| Highest previous amount | 59.38(71.18) | 59.60(73.05) | 58.96(67.27) |
| # of months since last donation | 13.01(12.08) | 13.01(12.09) | 13.00(12.07) |
| Already donated since 2005 | 0.52(0.50) | 0.52(0.50) | 0.52(0.50) |
| **Census Demographics** | | | |
| Female | 0.28(0.45) | 0.28(0.45) | 0.28(0.45) |
| Couple | 0.09(0.29) | 0.09(0.29) | 0.09(0.29) |
| **State and County** | | | |
| Red county | 0.40(0.49) | 0.41(0.49) | 0.40(0.49) |
| Red state | 0.51(0.50) | 0.51(0.50) | 0.51(0.50) |
| **Household Information** | | | |
| Average household size | 2.43(0.38) | 2.43(0.38) | 2.43(0.38) |
| Median household income | 54815.70(22027.32) | 54763.17(22074.82) | 54921.09(21932.01) |
| **Legal** | | | |
| Non litigation | 2.47(1.96) | 2.48(1.97) | 2.45(1.95) |
| Cases | 1.50(1.16) | 1.50(1.16) | 1.50(1.15) |

Table 1: The summary statistics of different subgroups of samples. Each entry in the table is the mean of the specified subgroup of all, treatment, or control samples. The number in the parentheses is the standard deviation.




(a) Dollars given without matching amount (x in log scale).

(b) Dollars given with matching amount (x in log scale).

(c) Donation amount change without matching amount.

Figure 2: The eCDF of dollars given without matching amount, dollars given with matching amount, and donation amount change without matching amount. The x-axis for 2a and 2b is in $\log_1 0$ scale since most of the values are 0, while the x-axis is at normal scale for 2c.

| Outcomes | Difference | Wald test | Permutation test |
|---|---|---|---|
| Response Rate | 0.0042 | 0.0013** | 0.0008*** |
| Dollars Given Not Including Match | 0.1536 | 0.0551* | 0.0300** |
| Dollars Given Including Match | 2.0876 | 0.0000*** | 0.0000*** |
| Amount Change | 6.3306 | 0.6374 | 0.2253 |

Table 2: Multiple outcomes case. We examine whether matching grants encourages receivers to donate. The listed values are $p$-values of the tests. *, **, and *** indicate that the corresponding p-value less than 10%. 5%, and 1% respectively.

where $D$ is the set of treatment status, $Z$ is the set of all subgroups, and $K$ is the set of outcome indices. The null hypotheses is then:

$$H_s : \mathbb{E}_Q[Y_{i,k}(d) - Y_{i,k}(d')|Z_i = z] = 0, \tag{4}$$

where $Y_{i,k}(d)$ is the $k$-th outcome of unit $i$ when the treatment $d$ is applied.

### 4.1 Multiple outcomes case

First we examine whether matching grants encourages receivers to donate. The set of null hypotheses in this case is indexed by the set:

$$S_1 = \{(d, d', z, k),\ d, d' \in D_1, z \in Z_1, k \in K_1\}, \tag{5}$$

where $D_1 = \{0, 1\}$ means we compare treatment group and control group, $Z_1 = \{0\}$ since we do consider subgroups in this case, and $K_1 = \{0, 1, 2, 3\}$ since we are examining all four outcomes. The total number of hypotheses is $|S_1| = 4$. The hypothesis tested here follows [3], yet the test procedure is different.

Table 2 lists the $p$-values of the tests. It suggests that the treatment influences the response rate, and dollars given with/without matching, while the amount change is not influenced.

### 4.2 Multiple subgroups case

This parts we examines whether the treatment influences the response rate for people residing in different states and counties. Similarly, the indices set $S_2$ is characterized by $D_2 = \{0, 1\}$, $Z_2 = \{0, 1, 2, 3\}$ representing the four combinations of red/blue states and red/blue counties, and $K_2 = \{0\}$ since we are examining the response rate. The total number of hypotheses is $|S_2| = 4$. The hypothesis tested here follows [3], yet the test procedure is different.

The test results in table 3 suggest that matching grants are working effectively for red states, while not effective in blue states. This is consistent with the conclusion listed in section 2.

| Subgroups | Difference | Wald test | Permutation test |
|---|---|---|---|
| Red County in a Red State | 0.0095 | 0.0000*** | 0.0000*** |
| Blue County in a Red State | 0.0070 | 0.0386** | 0.0192** |
| Red County in a Blue State | 0.0016 | 0.9935 | 0.4787 |
| Blue County in a Blue State | 0.0000 | 0.4899 | 0.2305 |

Table 3: Multiple subgroups case. We examine whether the treatment influences the response rate for people residing in different states and counties. The listed values are $p$-values of the tests. **, and *** indicate that the corresponding p-value less than 5% and 1% respectively.

| Outcomes | Difference | Wald test | Permutation test |
|---|---|---|---|
| Control vs 1:1 | -0.1234 | 0.2568 | 0.8770 |
| Control vs 2:1 | -0.2129 | 0.0508* | 0.9764 |
| Control vs 3:1 | -0.1245 | 0.2118 | 0.8929 |
| 1:1 vs 2:1 | -0.0895 | 0.4754 | 0.7610 |
| 1:1 vs 3:1 | -0.0011 | 0.9924 | 0.5039 |
| 2:1 vs 3:1 | 0.0883 | 0.4523 | 0.2267 |

Table 4: Multiple treatments case. whether the different matching ratios affects the amount given without grant amount. The listed values are $p$-values of the tests. *, **, and *** indicate that the corresponding p-value less than 10%. 5%, and 1% respectively.

### 4.3 Multiple treatments case

This parts we examines whether the different matching ratios affects the amount given without grant amount. Similarly, the indices set $S_3$ is characterized by $D_3 = \{0, 1, 2, 3\}$, $Z_3 = \{0\}$ since no subgroup is considered, and $K_3\{1\}$ since we are examining the dollars given without matching amount. The total number of hypotheses is $|S_3| = 6$. The latter three null hypotheses are not reported in [2].

From the results in table 4, applying matching grants influences the response ratio, the dollars given with/without including match if we look at the Wald results, which is consistent with the results in 4.1. However, the permutation results are significantly different from the Wald tests, which is possibly due the unbalanced distribution of data, since most of the values are zeros.

For the latter three hypotheses, we notice that the matching ratios do not have a influence on the revenue per letter. This is somehow counter-intuitive than the rule of thumb many believes.

## 5 Next Steps

This section lists the next steps to complete before the final report.

1. Understand the FWER proposed in [3] and replicate the results.
2. Apply the procedure to investigate the data more and make new conclusions, such as the relationship between the household income and treatments.

## References


[1] Dean Karlan and John A List. Does price matter in charitable giving? evidence from a large-scale natural field experiment. *American Economic Review*, 97(5):1774–1793, 2007.

[2] Nahum Kiryati and Yuval Landau. Dataset growth in medical image analysis research. *Journal of imaging*, 7(8):155, 2021.

[3] John A List, Azeem M Shaikh, and Yang Xu. Multiple hypothesis testing in experimental economics. *Experimental Economics*, 22(4):773–793, 2019.